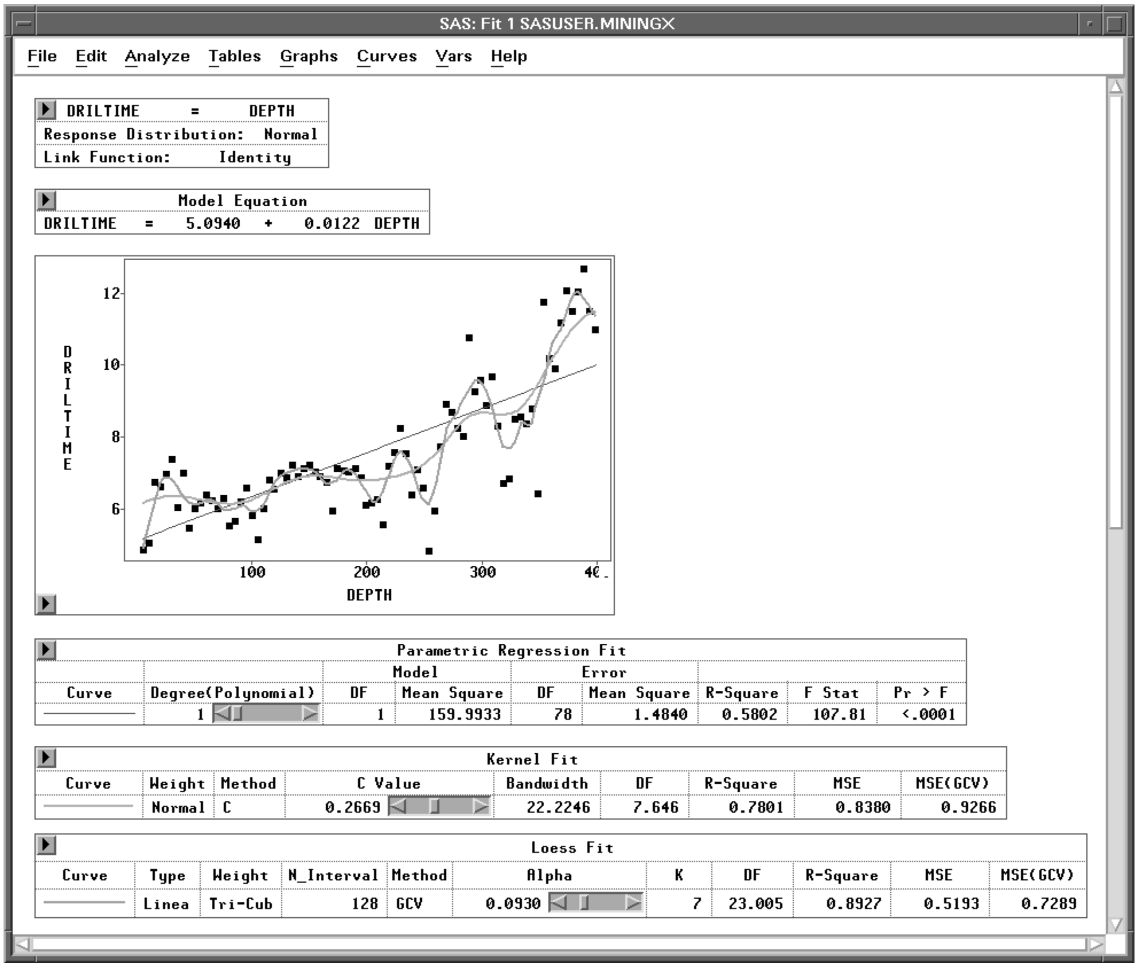
Chapter 13
Fitting Curves

Chapter Table of Contents

Chapter 13
Fitting Curves

You can use **Fit (Y X)** to fit curves when you have one **X** variable. Curve-fitting helps you identify trends and relationships in two-dimensional data. SAS/INSIGHT software offers both parametric and nonparametric methods to fit curves. You can generate confidence ellipses, fit parametric polynomials with confidence curves, and fit nonparametric curves using spline, kernel, and loess estimators.



DRILTIME = DEPTH
Response Distribution: Normal
Link Function: Identity

Model Equation
DRILTIME = 5.0940 + 0.0122 DEPTH

Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
—	1	1	159.9933	78	1.4840	0.5802	107.81	<.0001

Kernel Fit

Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
—	Normal	C	0.2669	22.2246	7.646	0.7801	0.8380	0.9266

Loess Fit

Curve	Type	Weight	N_Interval	Method	Alpha	K	DF	R-Square	MSE	MSE(GCV)
—	Linea	Tri-Cub	128	GCV	0.0930	7	23.005	0.8927	0.5193	0.7289

Figure 13.1. Fit Window with Several Curves

Parametric Regression Fits

Fitting a curve produces a visual display that reflects the systematic variation of the data. In this section, you will fit polynomial curves using a subset of the **MINING** data set described in Chapter 1, “Getting Started.”

⟹ **Open the MININGX data set.**

⟹ **Choose Analyze:Fit (Y X).**

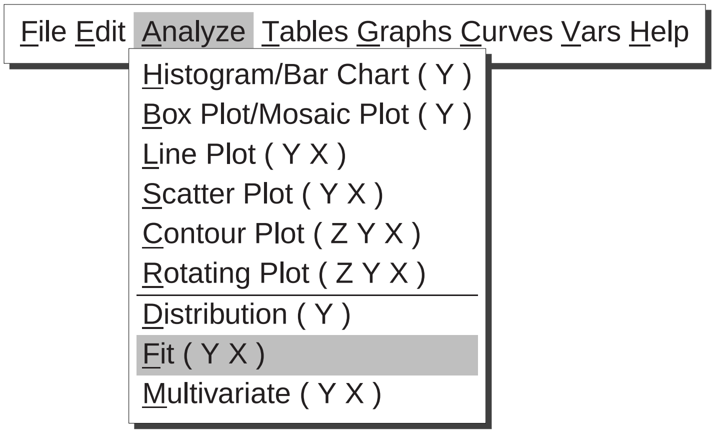


Figure 13.2. Analyze Menu

The fit variables dialog appears, as shown in Figure 13.3.

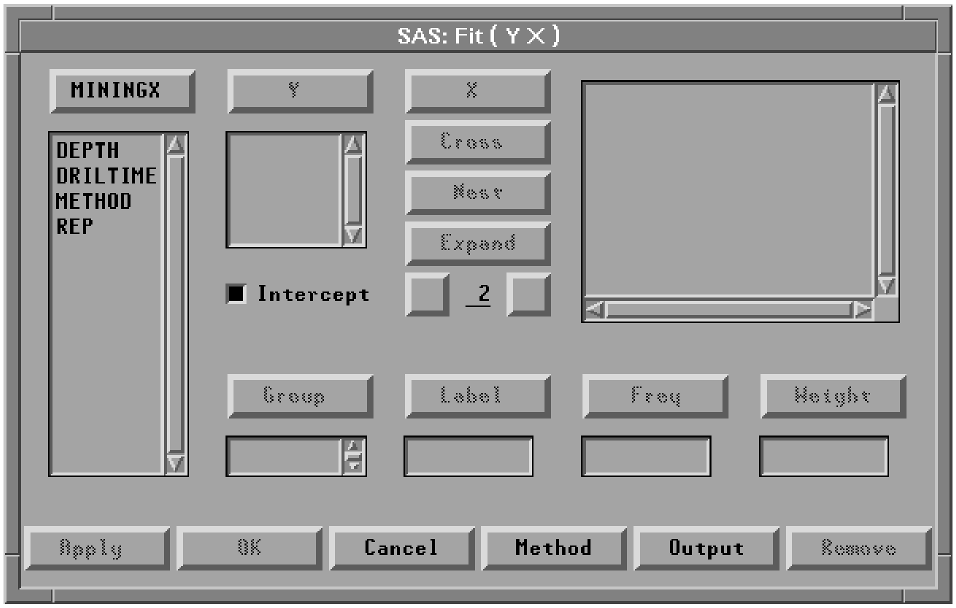
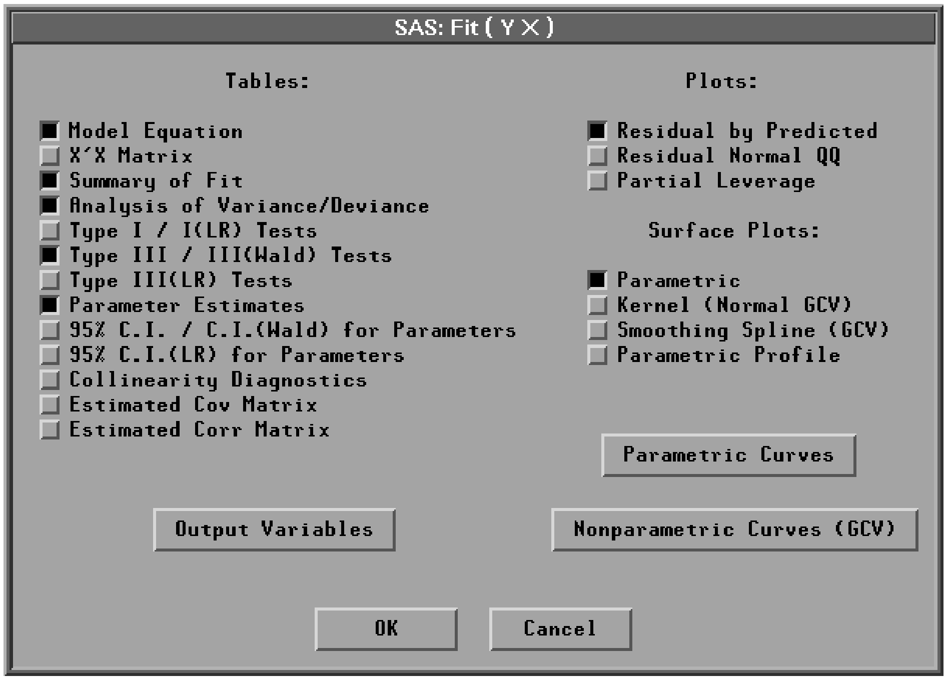


Figure 13.3. Fit Variables Dialog

⟹ **Select the variable DRILTIME, then click the Y button.**
DRILTIME appears in the **Y** variables list.

⟹ **Select the variable DEPTH, then click the X button.**
DEPTH appears in the **X** variables list.

⟹ **Click the Output button.**
The fit output options dialog, shown in Figure 13.4, appears on your display.



SAS: Fit (Y X)

Tables:

■ Model Equation
□ X'X Matrix
■ Summary of Fit
■ Analysis of Variance/Deviance
□ Type I / I(LR) Tests
■ Type III / III(Wald) Tests
□ Type III(LR) Tests
■ Parameter Estimates
□ 95% C.I. / C.I.(Wald) for Parameters
□ 95% C.I.(LR) for Parameters
□ Collinearity Diagnostics
□ Estimated Cov Matrix
□ Estimated Corr Matrix

Plots:

■ Residual by Predicted
□ Residual Normal QQ
□ Partial Leverage

Surface Plots:

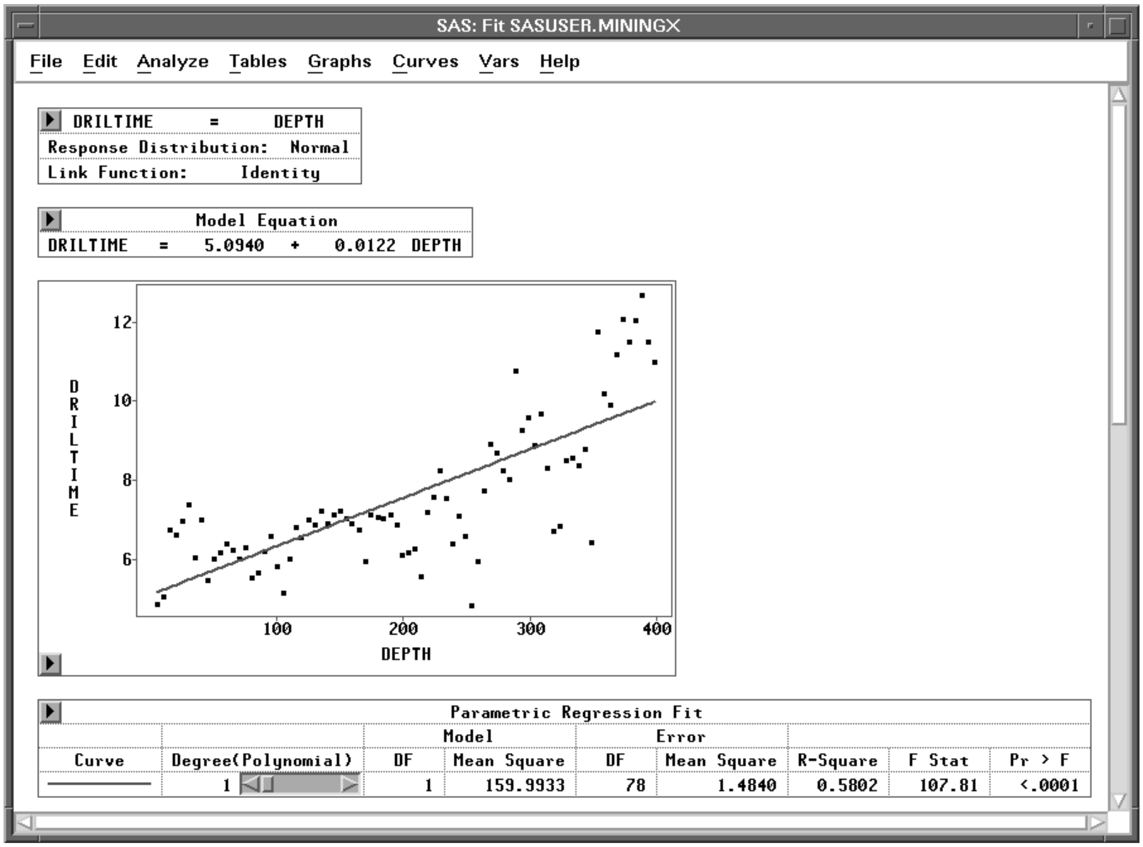
■ Parametric
□ Kernel (Normal GCV)
□ Smoothing Spline (GCV)
□ Parametric Profile

Figure 13.4. Fit Output Options Dialog

In the output options dialog, you specify which curves and tables will appear in the fit window. The default curve is a polynomial of degree one, that is, a line. The options set by default in this dialog are appropriate aids to a careful modeling of the data. They are not needed here where the purpose is to produce a visual display that reflects the trend of the data.

⟹ **Turn off all check boxes by clicking on any that are highlighted.**

⟹ **Click the OK button in all dialogs.**
A fit window appears, as shown in Figure 13.5.



DRILTIME = DEPTH
Response Distribution: Normal
Link Function: Identity

Model Equation
DRILTIME = 5.0940 + 0.0122 DEPTH

Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
———	1	1	159.9933	78	1.4840	0.5802	107.81	<.0001

Figure 13.5. Fit Window with Line

The fit window contains a plot of **DRILTIME** by **DEPTH** along with a table summarizing the fit. A simple regression line is superimposed on the plot; it follows the *linear* trend of the data. Notice, though, that the plot shows curvature that a straight line cannot follow.

First examine the **Parametric Regression Fit** table corresponding to these data. The **R-Square** value is **0.5802**, which means that 58% of the variation in drilling times is explained by **DEPTH**. The rest of this table contains statistics pertinent to hypothesis testing, and they are discussed in Chapter 14, “Multiple Regression.”

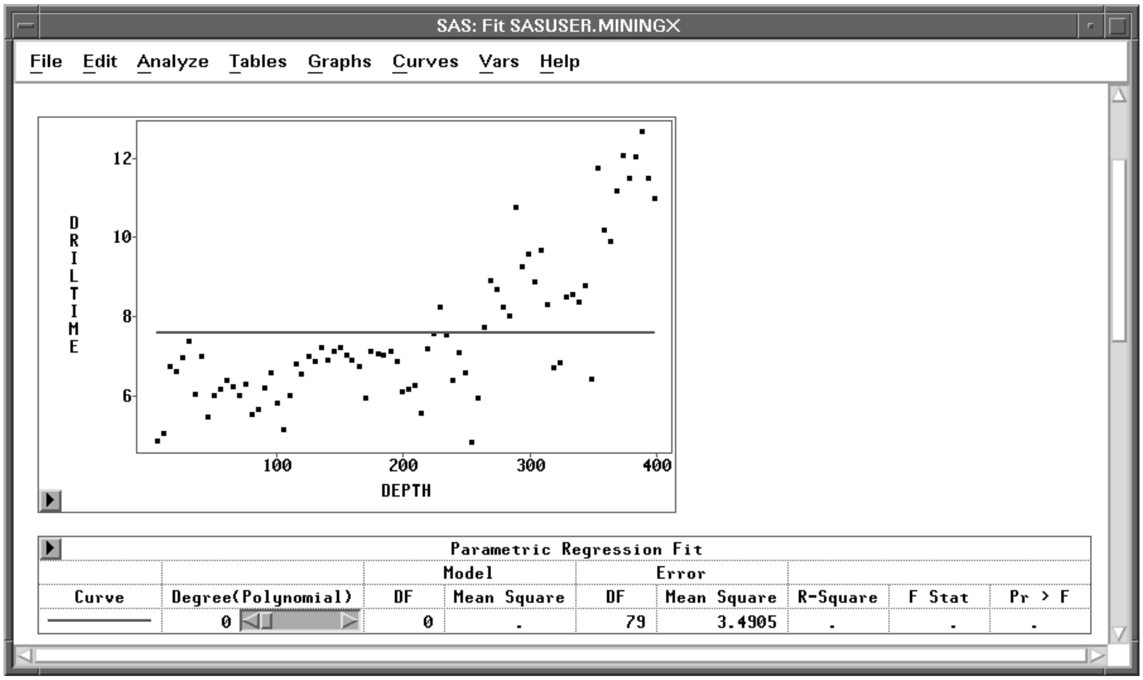
Changing the Polynomial Degree

Examine the **Parametric Regression Fit** table in Figure 13.6. Note that next to the polynomial degree is a slider that enables you to change the degree of polynomial fit to try to account for the curvature in the plot not explained by the straight line.

You can use the slider in three ways to adjust curves:

- click the arrow buttons
- click within the slider
- drag within the slider

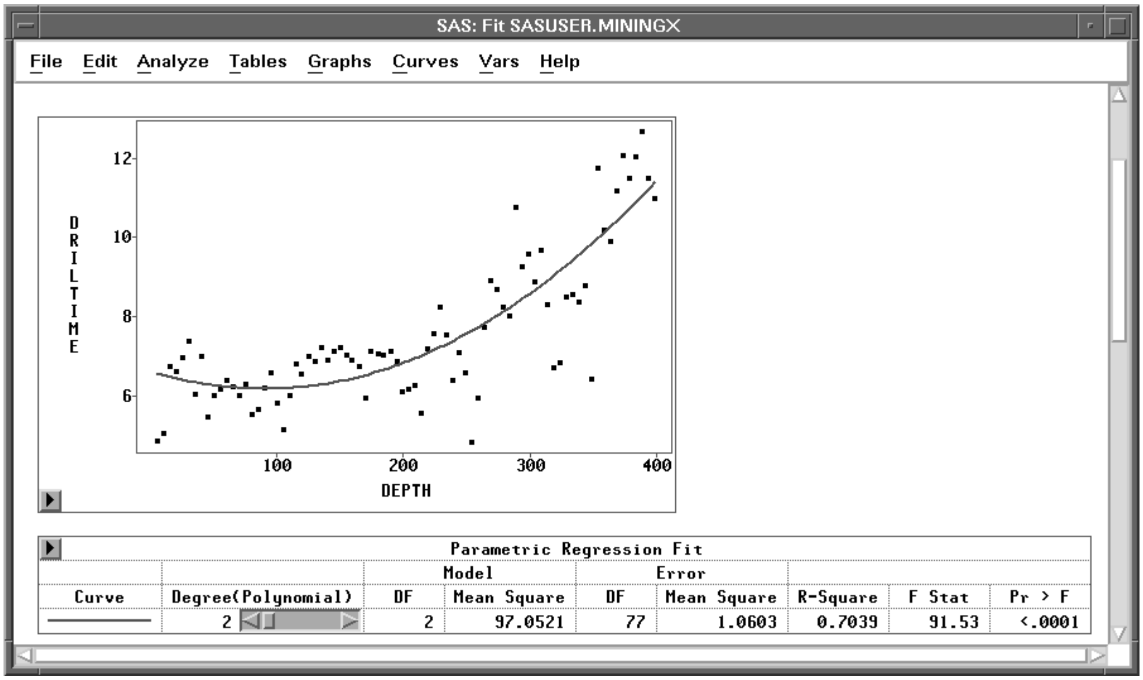
⟹ **Click the left arrow button in the slider.**
This decreases the degree of the polynomial to zero. A zero-degree polynomial fit is just a mean line.



Parametric Regression Fit								
		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
────	0	0	.	79	3.4905	.	.	.

Figure 13.6. Fit Window with Mean Line

⟹ **Click twice on the right arrow button in the slider.**
This increases the polynomial degree to **2**, a quadratic fit, as shown in Figure 13.7. The quadratic fit does a much better job accounting for the curvature in the plot. Note also that the **R-Square** value for the quadratic polynomial has increased to over 70%. You can fit successively higher-degree polynomials that continue to increase the **R-Square** value; but beyond a certain degree, small increases in **R-Square** do not compensate for the intuitive appeal in fitting a low degree polynomial.



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
—	2	2	97.0521	77	1.0603	0.7039	91.53	<.0001

Figure 13.7. Quadratic Fit

⟹ **Click within the slider, just to the right of the slider control.**
This moves the slider control to the position where you click. The polynomial degree is set to a value proportional to the slider position. On most personal computers, clicking within the slider is the fastest way to adjust a curve.

⟹ **Drag the slider control left and right.**
When you drag the slider, its speed depends on the number of data points, the type of curve, and the speed of your host. Depending on your host, you may be able to improve the speed of the dynamic graphics with an alternate drawing algorithm. To try this, choose **Edit:Windows:Graph Options**, and set the **Fast Draw** option.

† **Note:** The **Degree(Polynomial)** is the degree being specified in the polynomial fit, and the **Model DF** is the polynomial degree actually fitted.

To avoid unnecessary computation, the maximum degree that can be actually fitted is not calculated, and the maximum **Degree(Polynomial)** in the slider is set to be the number of unique **X** variable values minus 1. When a polynomial term for the **X** variable in the specified polynomial fit is a linear combination of its lower polynomial terms, the **Degree(Polynomial)** will be greater than the **Model DF**; that is, the degree actually fitted is less than the degree specified in these cases..

Adding Curves

You can add curves to a scatter plot in the fit window in two ways. You can choose from the **Curves** menu or you can select **Edit:Windows:Renew** to reset the fit output options. When you add a curve from the **Curves** menu, SAS/INSIGHT adds either a new table entry or a whole new table that contains a summary of the new curve fit. Suppose you want to compare polynomial fits of different degree directly on the scatter plot. Begin by adding a second polynomial fit to the plot.

⟹ **Choose Curves:Polynomial.**

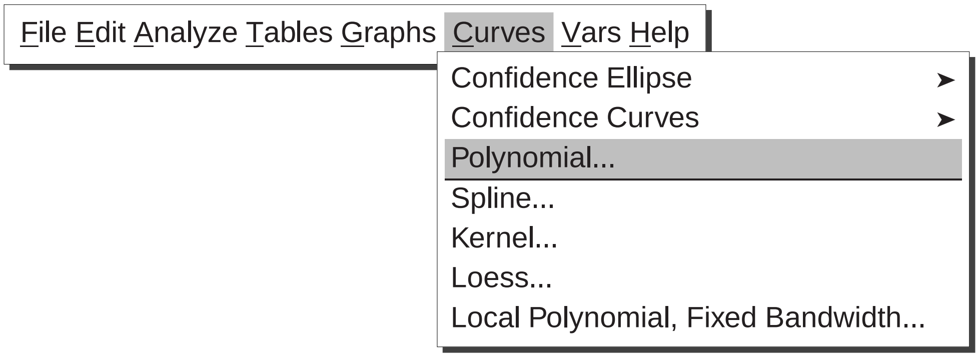
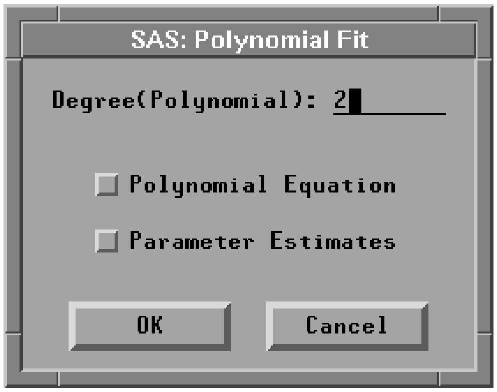

File Edit Analyze Tables Graphs Curves Vars Help
Confidence Ellipse ➤
Confidence Curves ➤
Polynomial...
Spline...
Kernel...
Loess...
Local Polynomial, Fixed Bandwidth...

Figure 13.8. Curves Menu

This displays the polynomial fit dialog shown in Figure 13.9.



Degree(Polynomial): 2

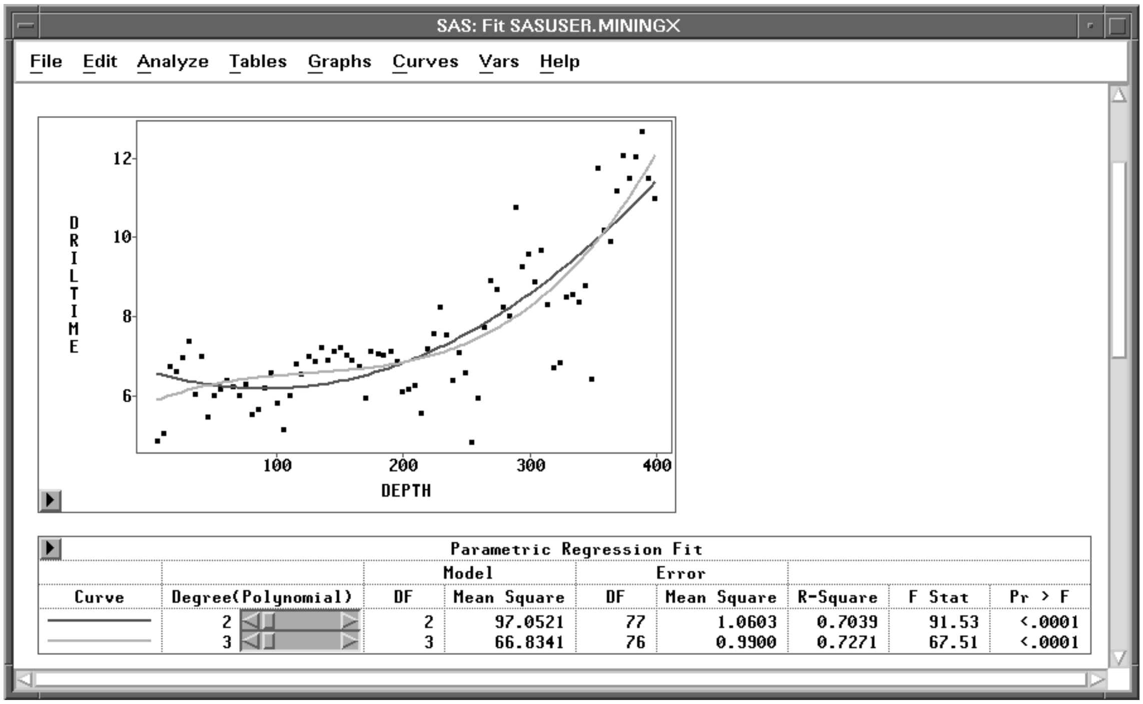
□ Polynomial Equation

□ Parameter Estimates

Figure 13.9. Polynomial Fit Dialog

⟹ **Set the degree for the new polynomial to 3 and click OK.**
This adds a cubic polynomial fit to the scatter plot, as shown in Figure 13.10.

Now you have two polynomial fits in the window. Note that an entry for the cubic polynomial has been added to the **Parametric Regression Fit** table. Each entry in the table has its own slider so that you can adjust the degree of either polynomial to compare any pair of fits.



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
───	2	2	97.0521	77	1.0603	0.7039	91.53	<.0001
───	3	3	66.8341	76	0.9900	0.7271	67.51	<.0001

Figure 13.10. Fit Window with Two Polynomial Fits

Line Colors, Patterns, and Widths

Notice in Figure 13.10 that it is difficult to distinguish the two polynomial curves. On color displays, curve colors are chosen by default to contrast with the window background color and with existing curves. Curves are always drawn as solid lines by default. You can set default curve widths with display options. You can use the **Tools** window to change any of these curve features.

⟹ **Choose Edit:Windows:Tools to display the tools window.**
The tools window displays a palette of colors, three line patterns, and five curve widths that you can choose for the selected curve, as shown in Figure 13.11

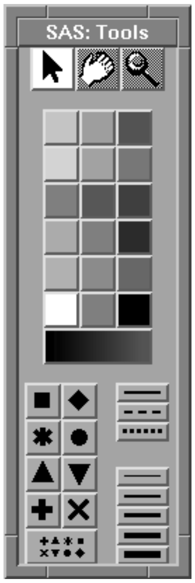
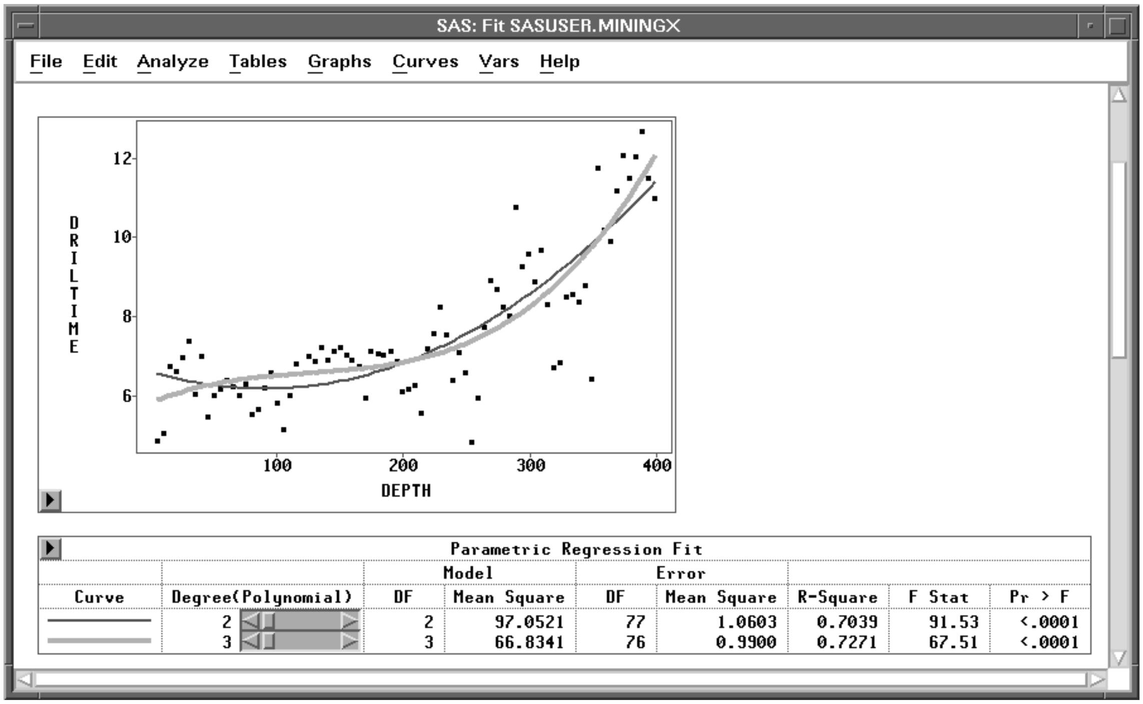


Figure 13.11. Tools Window

⟹ **Click on the cubic fit curve legend to select the curve.**

Clicking on either the legend or the curve highlights both the legend and the curve.

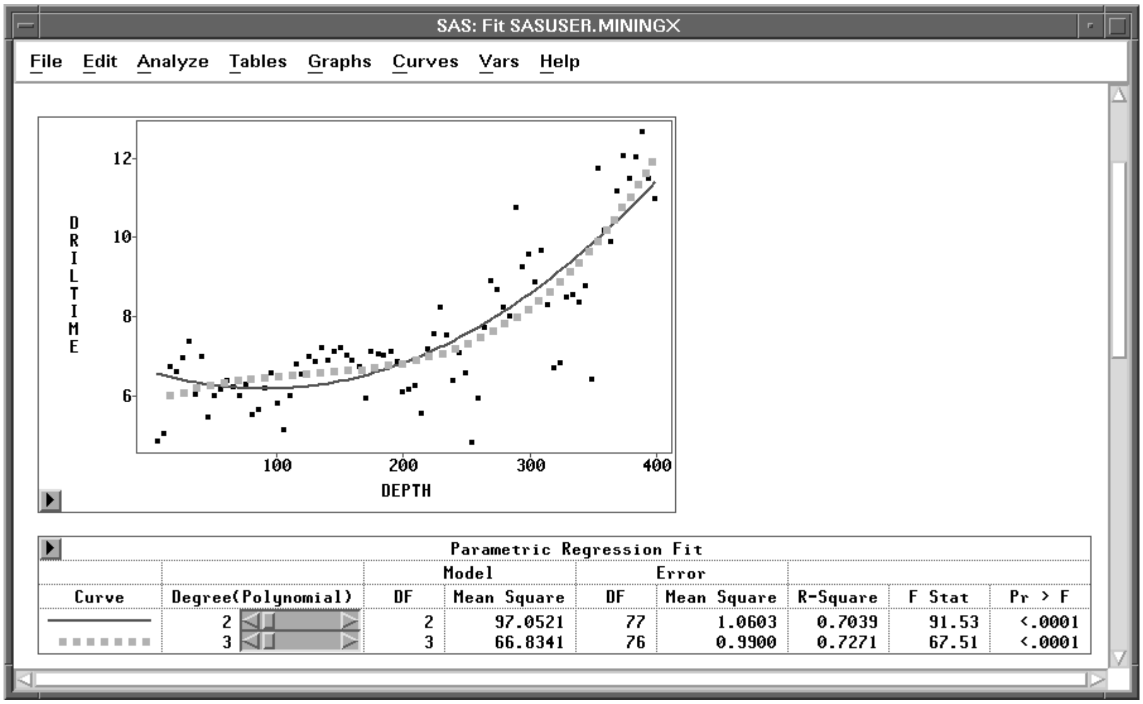


Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
—	2	2	97.0521	77	1.0603	0.7039	91.53	<.0001
—	3	3	66.8341	76	0.9900	0.7271	67.51	<.0001

Figure 13.12. Cubic Fit Curve Selected

⟹ **In the Tools window, click on the dotted line pattern.**
Again note that the legend in the table matches the new curve pattern.



Parametric Regression Fit

		Model		Error				
Curve	Degree(Polynomial)	DF	Mean Square	DF	Mean Square	R-Square	F Stat	Pr > F
────	2	2	97.0521	77	1.0603	0.7039	91.53	<.0001
■ ■ ■ ■ ■ ■ ■	3	3	66.8341	76	0.9900	0.7271	67.51	<.0001

Figure 13.13. New Pattern for Cubic Fit

⟹ **Click in any blank area of the fit window to deselect the curve.**
You can select a curve again and try various colors, patterns, or widths.

⟹ **Select the Parametric Regression Fit table.**

⟹ **Choose Edit:Delete.**
The selected parametric regression fit table and its associated curves are deleted from the window.

Nonparametric Fits

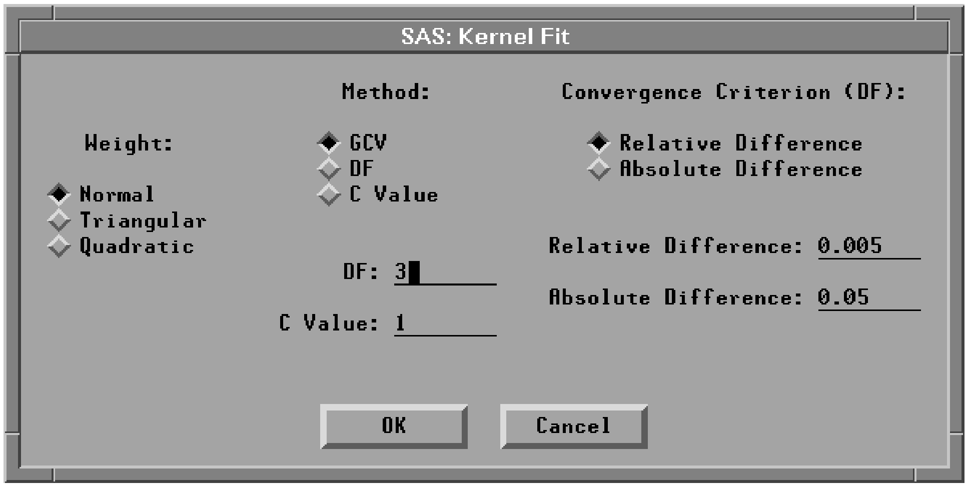
SAS/INSIGHT software provides nonparametric curve-fitting estimates from smoothing spline, kernel, loess, and fixed bandwidth local polynomial estimators that are alternatives to fitting polynomials. Because nonparametric methods allow more flexibility for the functional dependence of Y on X than a typical parametric model does, nonparametric methods are well suited for situations where little is known about the process under study.

To carry out a nonparametric regression, you need first to determine the smoothness of the fit. With SAS/INSIGHT software, you can specify a particular value for a smoothing parameter, specify a particular degrees of freedom for a smoother, or request a default best fit. The data are then smoothed to estimate the regression curve. This is in contrast to the parametric regression where the degree of the polynomial controls the complexity of the fit. For the polynomial, additional complexity can result in inappropriate global behavior. Nonparametric methods allow local use of additional complexity and thus are better tools to capture complex behavior than polynomials.

Normal Kernel Fit

To add a normal kernel estimate in the **MININGX** fit window from the preceding section, follow these steps.

⟹ **Choose Curves:Kernel.**
This displays the kernel fit dialog, as shown in Figure 13.14.



SAS: Kernel Fit

Weight:
● Normal
○ Triangular
○ Quadratic

Method:
● GCV
○ DF
○ C Value

DF: 3

C Value: 1

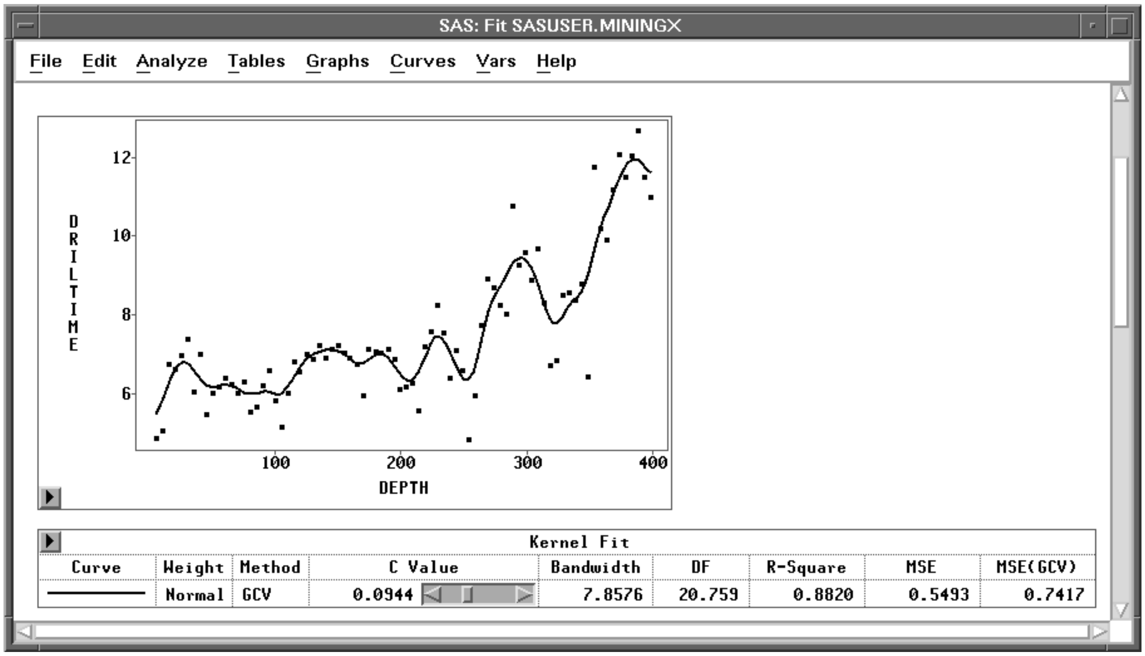
Convergence Criterion (DF):
● Relative Difference
○ Absolute Difference

Relative Difference: 0.005

Absolute Difference: 0.05

Figure 13.14. Kernel Fit Dialog

⟹ **Click on OK in the dialog to display the kernel fit, as shown in Figure 13.15.**



Kernel Fit

Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
———	Normal	GCV	0.0944	7.8576	20.759	0.8820	0.5493	0.7417

Figure 13.15. Normal Kernel Fit

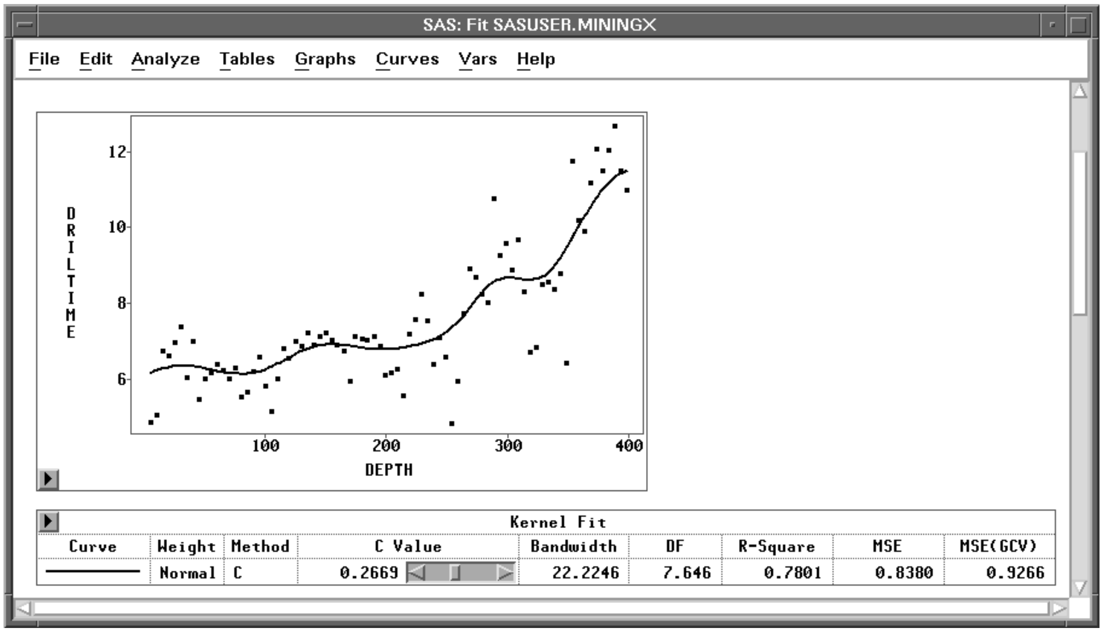
By default, the optimal kernel smoothness is estimated based on mean square error using *generalized cross validation* (GCV). Cross validation leaves out points (x_i, y_i) one at a time and computes the kernel regression at x_i based on the remaining *n*-1 observations. Generalized cross validation is a weighted version of cross validation and is easier to compute. This estimation is carried out for a number of different values of the smoothing parameter, and the value that minimizes the estimated mean square error is selected (Hastie and Tibshirani 1990). This technique is described in detail in Chapter 39, "Fit Analyses." Note that in Figure 13.15, the **Kernel Fit** table shows the **Method** as **GCV**.

You can change the degree of smoothness by using the slider in the **Kernel Fit** table to change the value of *c*. Higher values of *c* result in smoother curves closer to a straight line; smaller values produce more flexible curves. It is often necessary to experiment with several values before finding one that fits your data well. See Chapter 39, "Fit Analyses," for detailed information about kernels and the *c* parameter. Note that if you use the slider to change the value of *c*, the **Method** entry also changes.

The **Kernel Fit** table contains several statistics for comparing the kernel fit to other fits. The table contains the bandwidth or smoothing parameter of the kernel that corresponds to the value of *c*. The column labeled **DF** gives the approximate degrees of freedom for the kernel fit. Smoother curves have fewer degrees of freedom and result in lower values of R^2 and possibly higher values of mean square error. **R-square** measures the proportion of the total variation accounted for by the kernel fit. **MSE(GCV)** is an estimate of the mean square error using generalized cross validation. These statistics are also discussed in Chapter 39, "Fit Analyses."

This kernel tracks the data fairly well. The fit requires 20.759 degrees of freedom, indicating that the model may still be under-smoothed. The generalized cross validation method often results in under-smoothed fits, particularly with small data sets (Hastie and Tibshirani 1990). In this case, the data were collected from a single drilling hole, and this can lead to spurious cyclical patterns in the data caused by autocorrelation. The curve may be tracking these cycles. A smoother fit is probably desirable.

⟹ **Click three times on the right arrow in the slider.**
This results in a smoother kernel fit, as shown in Figure 13.16.



Kernel Fit

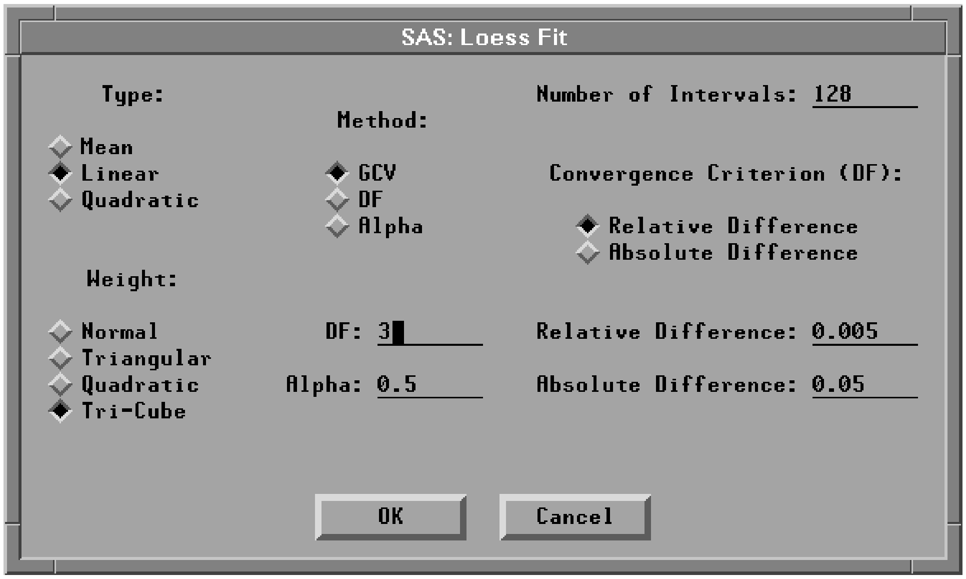
Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
———	Normal	C	0.2669	22.2246	7.646	0.7801	0.8380	0.9266

Figure 13.16. Normal Kernel Fit Made Smoother

Loess Smoothing

Loess smoothing is a curve-fitting technique based on local regression (Cleveland 1993). To fit a loess curve to the mining data, follow these steps:

⟹ **Choose Curves:Loess to display the loess fit dialog.**



SAS: Loess Fit

Type:
◇ Mean
◆ Linear
◇ Quadratic

Method:
◆ GCV
◇ DF
◇ Alpha

Number of Intervals: 128

Convergence Criterion (DF):
◆ Relative Difference
◇ Absolute Difference

Weight:
◇ Normal
◇ Triangular
◇ Quadratic
◆ Tri-Cube

DF: 3

Alpha: 0.5

Relative Difference: 0.005

Absolute Difference: 0.05

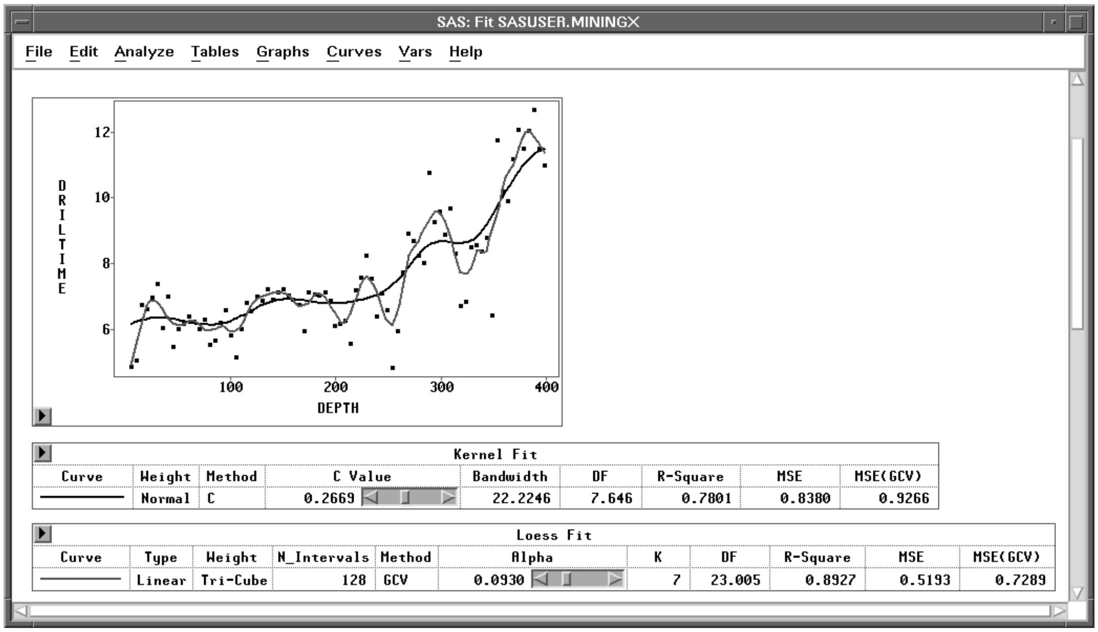
Figure 13.17. Loess Fit Dialog

⟹ **Click on OK in the dialog to display the loess fit, as shown in Figure 13.18.**
As with the kernel fit, the best fit for loess smoothing is determined by generalized cross validation (GCV). GCV and other aspects of curve-fitting are described in Chapter 39, "Fit Analyses."

You can also output predicted values from fitted curves. To output predicted values from the preceding loess fit, do the following:

⟹ **Choose Vars:Predicted Curves:Loess.**
This displays the same loess fit dialog as shown in Figure 13.17.

⟹ **Click on OK in the dialog to output the predicted values from the loess fit.**
A new variable, **PL_DRILT**, should now be added to the data window.



Kernel Fit

Curve	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
—	Normal	C	0.2669	22.2246	7.646	0.7801	0.8380	0.9266

Loess Fit

Curve	Type	Weight	N_Intervals	Method	Alpha	K	DF	R-Square	MSE	MSE(GCV)
—	Linear	Tri-Cube	128	GCV	0.0930	7	23.005	0.8927	0.5193	0.7289

Figure 13.18. Loess Fit

You can use the slider control to adjust the loess curve just as with other curves. For loess, the slider controls the α value for the fit. The greater the α value, the smoother the fit.

On rare occasions, you may want to fit a curve for α values outside the bounds of the slider. For loess and other curves, the bounds of the slider are chosen for best fit in most cases. If you need to fit a curve with unusual parameter values, you can specify these values in the curve dialog.

⊕ **Related Reading:** Fit Curves, Chapter 39.

References

Cleveland, W.S. (1993), *Visualizing Data*, Summit, New Jersey: Hobart Press.

Hastie, Y.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/ INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

The Institute is a private company devoted to the support and further development of its software and related services.